Analytical Statistics using RStudio

Edre MA, DrPH

2020-12-16

```
#=======================
#Analytical statistics
#R Biostat Workshop IIUM
#Edre MA, DrPH
#=======================

#objective 3: To determine the factors contributing to hypertension
#we want to know first what contributes to systolic hypertension

#Comparing numerical values: parametric

install.packages('readr', repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/raef/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'readr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\raef\AppData\Local\Temp\RtmpCaiI7n\downloaded_packages
```

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 3.6.3
```

```
healthstat <- read_csv("healthstatus6.csv")
```

```
##
## -- Column specification --------------------------------------------------
---------
## cols(
##   id = col_double(),
##   age = col_double(),
##   sex = col_character(),
##   exercise = col_character(),
##   smoking = col_character(),
##   wt = col_double(),
##   ht = col_double(),
##   sbp = col_double(),
##   dbp = col_double(),
##   hba1c = col_double(),
##   hcy = col_double(),
##   wt2 = col_double(),
##   wt3 = col_double(),
##   sbp2 = col_double(),
##   sbp3 = col_double(),
##   dbp2 = col_double(),
##   dbp3 = col_double()
## )
```

```
View(healthstat)

#if our IV is categorical with 2 categories
#example, we want to know if being male has any relationship with sbp
#independent sample t test

install.packages("car", , repos = "http://cran.us.r-project.org") #testing fo
r homogeneity of variance
```

```
## Installing package into 'C:/Users/raef/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'car' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\raef\AppData\Local\Temp\RtmpCaiI7n\downloaded_packages
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.3
```

```
## Loading required package: carData
```

```
leveneTest(sbp ~ sex, data = healthstat, center=mean)
```

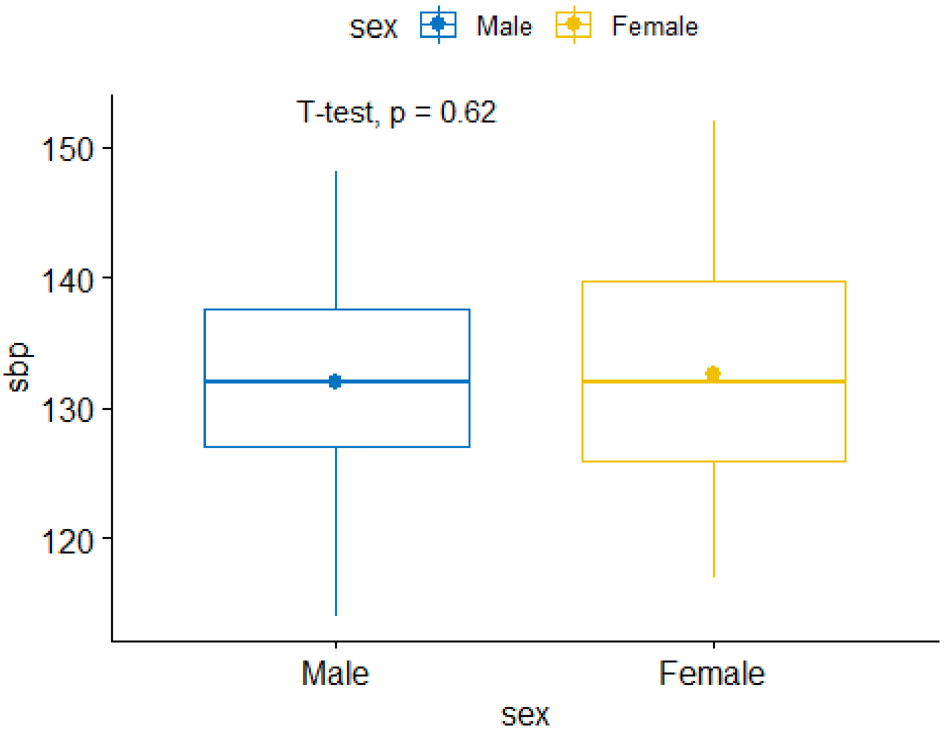
```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##        Df F value Pr(>F)
## group   1  0.5477 0.4604
##       151
```

```
t.test(sbp ~ sex, data = healthstat)
```

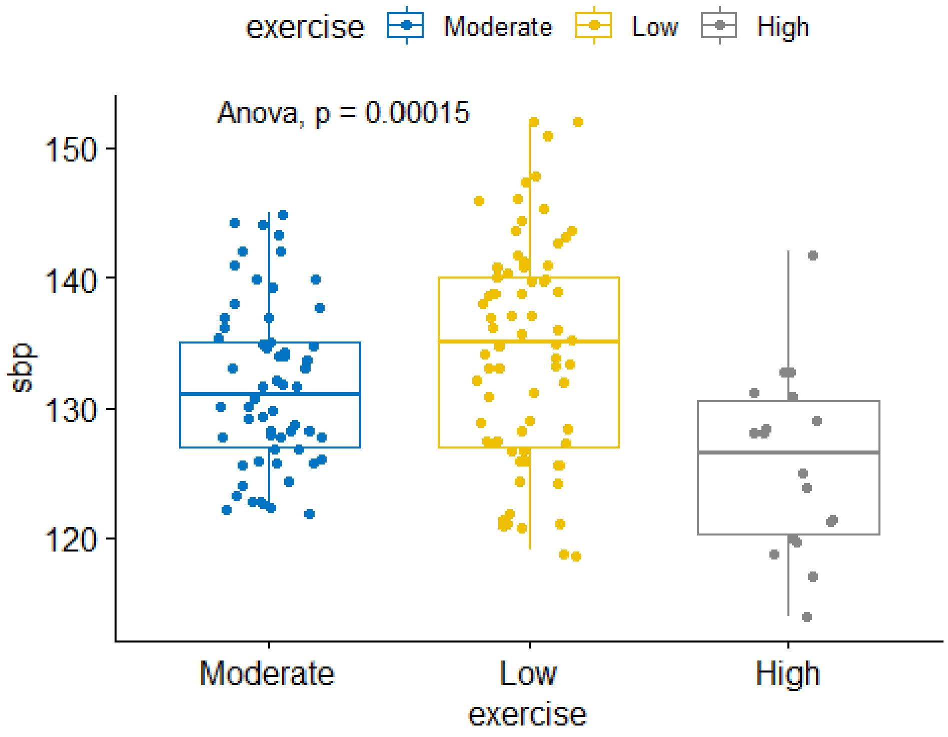
```
##
##  Welch Two Sample t-test
##
## data:  sbp by sex
## t = 0.4972, df = 141.8, p-value = 0.6198
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.928972  3.225357
## sample estimates:
```

```
## mean in group Female   mean in group Male 
##              132.6000             131.9518

#we want to visualize the comparison

library(ggpubr)

## Warning: package 'ggpubr' was built under R version 3.6.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.6.3

ggboxplot(healthstat, x = "sex", y = "sbp",
          color = "sex",
          palette = "jco",
          add = "mean") +
          stat_compare_means(method = "t.test")

## Warning: `fun.y` is deprecated. Use `fun` instead.

## Warning: `fun.ymin` is deprecated. Use `fun.min` instead.

## Warning: `fun.ymax` is deprecated. Use `fun.max` instead.
```



```
#now we know sex has no effect on sbp in our study
#we want to know now, does exercise have an effect (low,mod,high intensity)
#one way ANOVA
```

```r
library(psych)

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

## The following object is masked from 'package:car':
##
##     logit

describe.by(healthstat$sbp, healthstat$exercise)

## Warning: describe.by is deprecated. Please use the describeBy function

##
##  Descriptive statistics by group
## group: High
##    vars  n   mean   sd median trimmed  mad min max range skew kurtosis   s
e
## X1    1 18 125.78 6.97  126.5   125.5 8.15 114 142    28 0.35    -0.48 1.6
4
## ------------------------------------------------------------
## group: Low
##    vars  n   mean   sd median trimmed   mad min max range skew kurtosis se
## X1    1 74 134.24 8.57    135  134.15 10.38 119 152    33 0.03    -0.92  1
## ------------------------------------------------------------
## group: Moderate
##    vars  n   mean   sd median trimmed  mad min max range skew kurtosis  se
## X1    1 61 131.74 6.28    131  131.45 5.93 122 145    23 0.33    -0.84 0.8

one.way =aov(sbp ~ exercise, data = healthstat)
summary(one.way)

##              Df Sum Sq Mean Sq F value   Pr(>F)
## exercise      2   1064   532.0   9.324 0.000152 ***
## Residuals   150   8559    57.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ggboxplot(healthstat, x = "exercise", y = "sbp",
          color = "exercise",
          palette = "jco",
          add = "jitter") +
          stat_compare_means(method = "anova")
```



```
#yes, there is significant effect of exercise on sbp
#but, which pair comparison has most effect?

leveneTest(sbp ~ exercise, data = healthstat, center=mean)

## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = mean)
##        Df F value  Pr(>F)
## group   2  4.2458 0.01608 *
##       150
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#significant p value, thus equal variance not assumed
#Post Hoc test - Games Howell

install.packages("userfriendlyscience", , repos = "http://cran.us.r-project.o
rg")

## Installing package into 'C:/Users/raef/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)

## package 'userfriendlyscience' successfully unpacked and MD5 sums checked
##
```

```
## The downloaded binary packages are in
##  C:\Users\raef\AppData\Local\Temp\RtmpCaiI7n\downloaded_packages
```

```r
library(userfriendlyscience)
```

```
## Warning: package 'userfriendlyscience' was built under R version 3.6.3
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

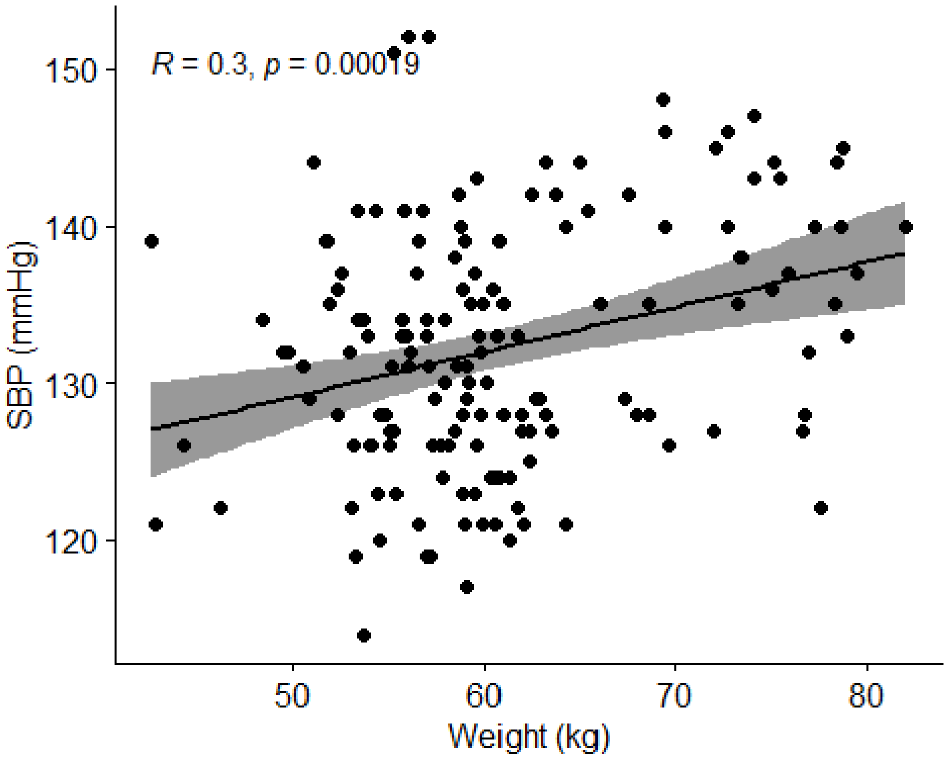
```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```r
oneway(healthstat$exercise, y = healthstat$sbp, posthoc = 'games-howell')
```

```
## Warning in oneway(healthstat$exercise, y = healthstat$sbp, posthoc = "game
s-
## howell"): ### Warning: the x variable (exercise) is not a factor! Converti
ng it
## myself - but note that variables in R have data types, and it's advisable
to set
## these adequately (use for example 'as.factor'; see '?as.factor' for help)!
```

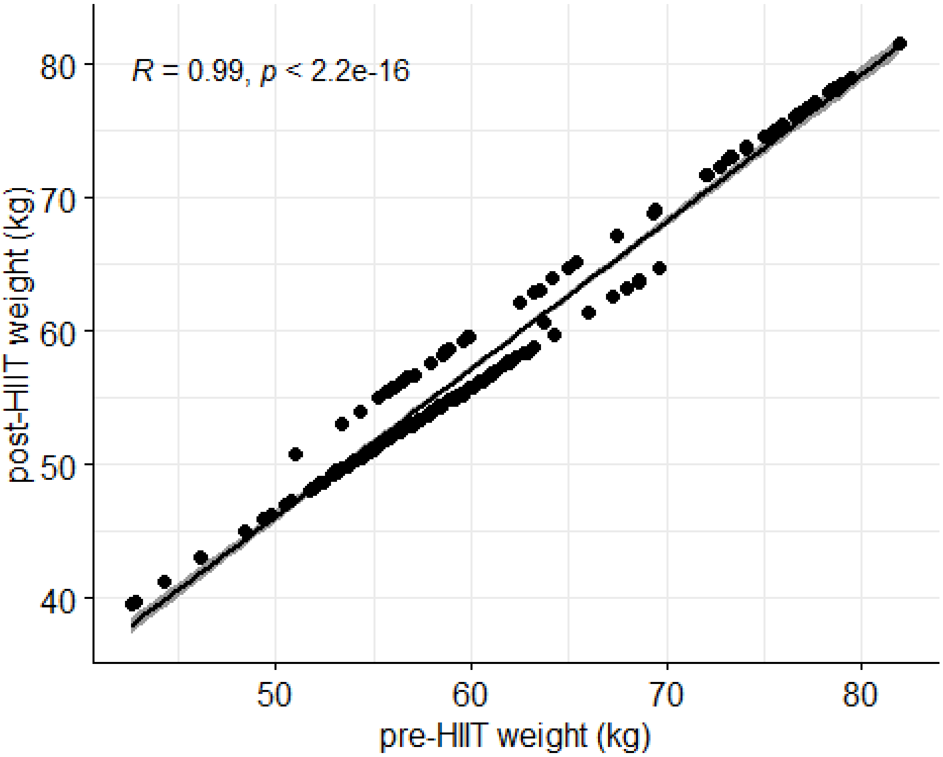
```
## ### Oneway Anova for y=sbp and x=exercise (groups: High, Low, Moderate)
```

```
## Registered S3 methods overwritten by 'ufs':
##   method                     from
##   grid.draw.ggProportionPlot userfriendlyscience
##   pander.associationMatrix   userfriendlyscience
##   pander.dataShape           userfriendlyscience
##   pander.descr               userfriendlyscience
##   pander.normalityAssessment userfriendlyscience
##   print.CramersV             userfriendlyscience
##   print.associationMatrix    userfriendlyscience
##   print.confIntOmegaSq       userfriendlyscience
##   print.confIntV             userfriendlyscience
##   print.dataShape            userfriendlyscience
##   print.descr                userfriendlyscience
##   print.ggProportionPlot     userfriendlyscience
##   print.meanConfInt          userfriendlyscience
##   print.multiVarFreq         userfriendlyscience
##   print.normalityAssessment  userfriendlyscience
##   print.regrInfluential      userfriendlyscience
##   print.scaleDiagnosis       userfriendlyscience
##   print.scaleStructure       userfriendlyscience
##   print.scatterMatrix        userfriendlyscience
## Omega squared: 95% CI = [.03; .2], point estimate = .1
## Eta Squared: 95% CI = [.04; .19], point estimate = .11
##
##                                       SS  Df     MS    F     p
## Between groups (error + effect) 1064.03   2 532.01 9.32 <.001
## Within groups (error only)      8558.54 150  57.06
##
##
## ### Post hoc test: games-howell
##
##                 diff ci.lo ci.hi    t     df     p
## Low-High       8.47  3.74 13.19 4.41  30.86 <.001
## Moderate-High  5.96  1.41 10.51 3.26  25.72  .009
## Moderate-Low  -2.51 -5.54  0.53 1.96 131.30  .127

#if equal variance assumed, use Tukey
tukey.one.way<-TukeyHSD(one.way)
tukey.one.way

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = sbp ~ exercise, data = healthstat)
##
## $exercise
##                     diff       lwr        upr     p adj
## Low-High       8.465465  3.766172 13.1647594 0.0001043
## Moderate-High  5.959927  1.163662 10.7561923 0.0105140
## Moderate-Low  -2.505538 -5.597805  0.5867282 0.1371181

#Exercise adds benefit in sbp reduction
#does it correlate with weight?

#pearson correlation coefficient test

cor.test(healthstat$wt,healthstat$sbp, method="pearson")

##
##  Pearson's product-moment correlation
##
## data:  healthstat$wt and healthstat$sbp
## t = 3.8267, df = 151, p-value = 0.0001897
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1455186 0.4354641
## sample estimates:
##       cor
## 0.2973312

ggscatter(healthstat, x = "wt", y = "sbp",
          add = "reg.line",
          conf.int = TRUE,
          cor.coef = TRUE,
          cor.method = "pearson",
          xlab = "Weight (kg)", ylab = "SBP (mmHg)")

## `geom_smooth()` using formula 'y ~ x'
```

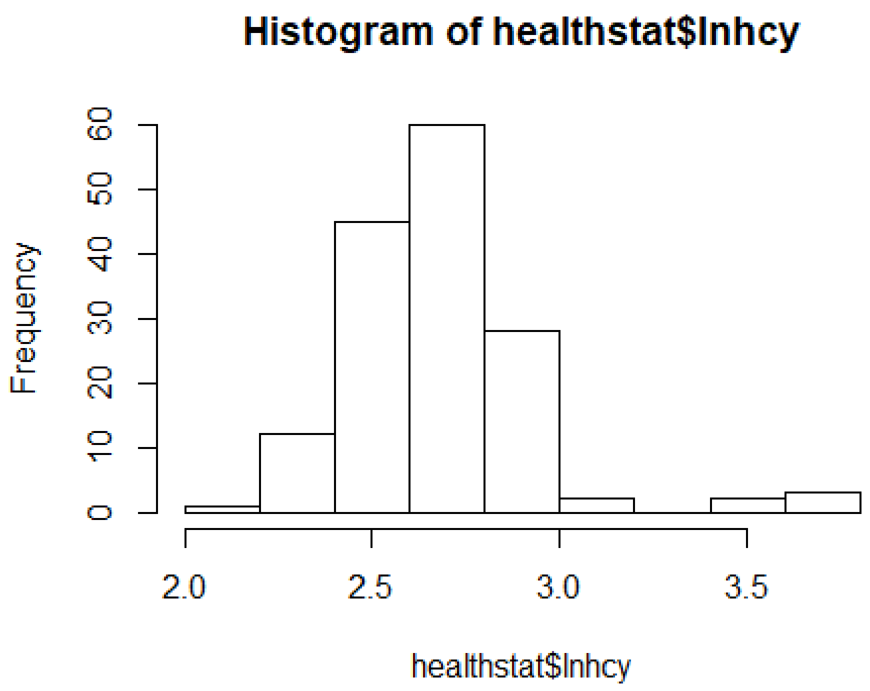


```
#yes, the heavier the person, the higher the sbp
#now you conducted a high-intensity interval training (HIIT) intervention
#you want to measure pre and post HIIT effect on weight
#paired t test

t.test(healthstat$wt, healthstat$wt2, paired=TRUE)

##
##  Paired t-test
##
## data:  healthstat$wt and healthstat$wt2
## t = 19.015, df = 152, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.445299 3.012348
## sample estimates:
## mean of the differences
##                2.728824
```

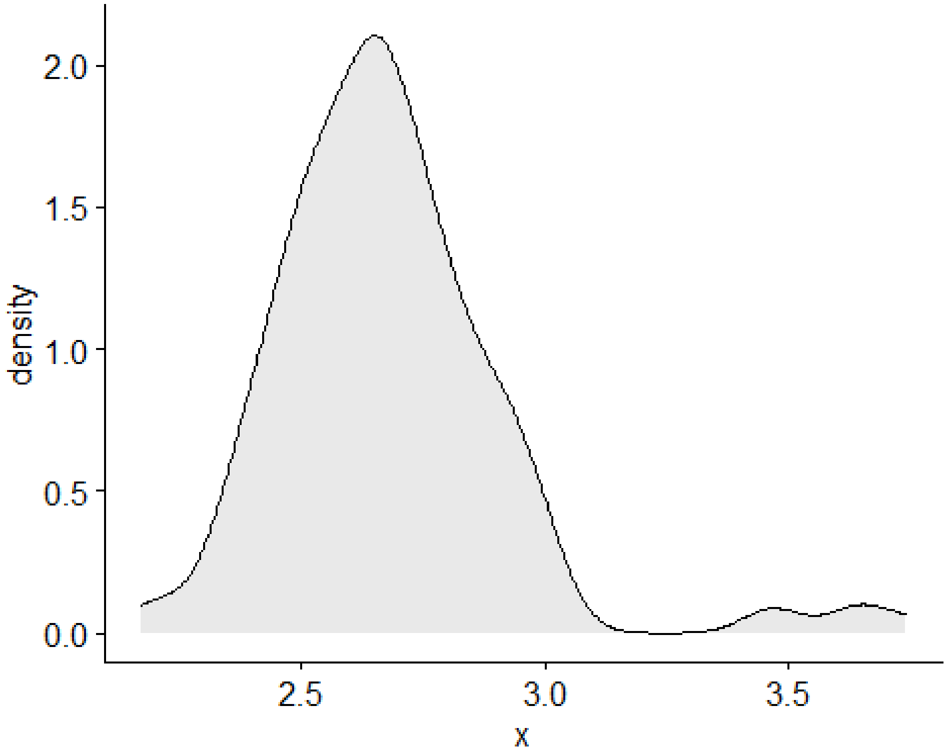
```
ggscatter(healthstat, x = "wt", y = "wt2",
          add = "reg.line",
          conf.int = TRUE,
          cor.coef = TRUE,
          cor.method = "pearson",
          xlab = "pre-HIIT weight (kg)", ylab = "post-HIIT weight (kg)") +
          grids(linetype = "solid")

## `geom_smooth()` using formula 'y ~ x'
```



```
#Now we know that sbp is affected by weight.
#exercise gives additional benefit to weight reduction
#you are now concerned with the exercise giving effect on cardiovascular heal
th
#homocysteine (hcy) relates to cardiovascular heath from literature
#what are the factors contributing to hcy level?

#comparing numerical values: non-parametric

#try transforming data into normal distribution by ln
healthstat$lnhcy= log(healthstat$hcy)
hist(healthstat$lnhcy)
```



```
ggdensity(healthstat$lnhcy, fill = "lightgray")
```



```
#still not normally distributed
#need to do non-parametric test

#female has higher or lower hcy level compared to male?
#mann whitney U test

install.packages("SmartEDA",repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/raef/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'SmartEDA' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\raef\AppData\Local\Temp\RtmpCaiI7n\downloaded_packages
```

```
library(SmartEDA)
```

```
## Warning: package 'SmartEDA' was built under R version 3.6.3
```

```
ExpCustomStat(healthstat,
              Cvar="sex",
              Nvar="hcy",
              stat=c("median","IQR"),
              gpby=TRUE,
              dcast=F)
```

```
##       sex Attribute median   IQR
## 1:   Male       hcy   14.3 3.650
## 2: Female       hcy   14.1 3.575
```

```
wilcox.test(hcy~sex, data=healthstat)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  hcy by sex
## W = 2658.5, p-value = 0.3675
## alternative hypothesis: true location shift is not equal to 0
```

```
#sex has no signififant relationship with hcy

#how about exercise intensity?
#kruskal wallis test


ExpCustomStat(healthstat,
              Cvar="exercise",
              Nvar="hcy",
              stat=c("median","IQR"),
              gpby=TRUE,
              dcast=F)
```

```
##    exercise Attribute median   IQR
## 1: Moderate       hcy  14.30 2.600
## 2:      Low       hcy  14.55 5.075
## 3:     High       hcy  12.35 2.350
```

```
kruskal.test(hcy ~ exercise, data = healthstat) #if significant, proceed with pairwise comparison
```

```
## 
##  Kruskal-Wallis rank sum test
## 
## data:  hcy by exercise
## Kruskal-Wallis chi-squared = 11.436, df = 2, p-value = 0.003286
```

```
pairwise.wilcox.test(healthstat$hcy, healthstat$exercise,p.adjust.method = "BH")
```

```
## 
##  Pairwise comparisons using Wilcoxon rank sum test 
## 
## data:  healthstat$hcy and healthstat$exercise 
## 
##          High  Low  
## Low      0.005 -    
## Moderate 0.014 0.133
## 
## P value adjustment method: BH
```

```
#high intensity exercise significantly gives lower HCY compared to low/mod
#you proceed in continuing the HIIT intervention as it gives benefit to both sbp and hcy

#measure effectiveness again on weight reduction
#wilcoxon signed rank test

wilcox.test(healthstat$wt,healthstat$wt2,paired=TRUE)
```

```
## 
##  Wilcoxon signed rank test with continuity correction
## 
## data:  healthstat$wt and healthstat$wt2
## V = 11781, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
#you notice some of your respondents are diabetic
#worry that your intervention gives more harm than good
#finding relationship between hba1c and both sbp/hcy
#spearman correlation coefficient test
```

```
cor.test(healthstat$hba1c,healthstat$sbp, method="spearman")
```

```
## Warning in cor.test.default(healthstat$hba1c, healthstat$sbp, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  healthstat$hba1c and healthstat$sbp
## S = 416312, p-value = 0.0001441
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3025482
```

```
cor.test(healthstat$hba1c,healthstat$hcy, method="spearman")
```

```
## Warning in cor.test.default(healthstat$hba1c, healthstat$hcy, method =
## "spearman"): Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  healthstat$hba1c and healthstat$hcy
## S = 515097, p-value = 0.09116
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1370514
```

```
#you conclude that only sbp has a significant correlation with hba1c
#in future, you would prioritize giving HITT intervention to hypertensive pat
ients

#now, you are focused back to your objective 3
#factors contributing to hypertension (hpt)

#comparing categorical variables

install.packages("dplyr",repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/raef/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'dplyr' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'dplyr'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying C:
## \Users\raef\Documents\R\win-library\3.6\00LOCK\dplyr\libs\x64\dplyr.dll to
C:
```

```
## \Users\raef\Documents\R\win-library\3.6\dplyr\libs\x64\dplyr.dll: Permissi
on
## denied

## Warning: restored 'dplyr'

##
## The downloaded binary packages are in
##  C:\Users\raef\AppData\Local\Temp\RtmpCaiI7n\downloaded_packages
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
healthstatcat<-healthstat %>%
  mutate(hpt=if_else(healthstat$sbp<140 & healthstat$dbp<90,'normal','high'))
View(healthstatcat)

#smoking has a relationship with hpt?
#chi square test

chisq.test(healthstatcat$hpt,healthstatcat$smoking,correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  healthstatcat$hpt and healthstatcat$smoking
## X-squared = 15.607, df = 1, p-value = 7.797e-05
```

```
chisq.test(healthstatcat$hpt,healthstatcat$smoking)$observed
```

```
##                   healthstatcat$smoking
## healthstatcat$hpt No Yes
##            high   23  36
##            normal 67  27
```

```
#yes, smoking is significantly related to hpt. More smokers are hypertensive

#how about BMI status and hpt?

healthstatcatbmi<- healthstatcat %>%
  mutate(height_m = ht / 100,bmi = wt / (height_m^2))
View(healthstatcatbmi)
healthstatcatbmi$bmistatus<- cut(healthstatcatbmi$bmi,
                                 breaks=c(-Inf, 18.49999, 24.9999, 29.9999, I
nf),
                                 labels=c("underweight", "normal", "overweigh
t", "obese"))

#fisher's exact test (used when more than 20% celss with expected count less
than 5)
chisq.test(healthstatcatbmi$hpt,healthstatcatbmi$bmistatus)$expected

## Warning in chisq.test(healthstatcatbmi$hpt, healthstatcatbmi$bmistatus): C
hi-
## squared approximation may be incorrect

##                       healthstatcatbmi$bmistatus
## healthstatcatbmi$hpt underweight   normal overweight     obese
##                 high      2.313725 28.92157    18.5098  9.254902
##                 normal    3.686275 46.07843    29.4902 14.745098

fisher.test(healthstatcatbmi$hpt,healthstatcatbmi$bmistatus)

##
##  Fisher's Exact Test for Count Data
##
## data:  healthstatcatbmi$hpt and healthstatcatbmi$bmistatus
## p-value = 1.205e-05
## alternative hypothesis: two.sided

#significant relationship between hpt and bmi status

#reporting your findings in table form

#package needed
#"sjPlot"
#"apaTables"

install.packages("sjPlot",repos = "http://cran.us.r-project.org")

## Installing package into 'C:/Users/raef/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)

## package 'sjPlot' successfully unpacked and MD5 sums checked
##
```

```
## The downloaded binary packages are in
##  C:\Users\raef\AppData\Local\Temp\RtmpCaiI7n\downloaded_packages
```

```r
library(sjPlot)
```

```
## Warning: package 'sjPlot' was built under R version 3.6.3
```

```
## Install package "strengejacke" from GitHub (`devtools::install_github("str
engejacke/strengejacke")`) to load all sj-packages at once!
```

```r
install.packages("apaTables",repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/raef/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'apaTables' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\raef\AppData\Local\Temp\RtmpCaiI7n\downloaded_packages
```

```r
library(apaTables)
```

```
## Warning: package 'apaTables' was built under R version 3.6.3
```

```r
#table created in word file in your directory!
sjt.xtab(healthstatcatbmi$smoking, healthstatcatbmi$hpt, file = "sjt_continge
ncy.doc")
```

smoking
hpt
Total
high
normal
No
23
67
90
Yes
36
27
63
Total
59
94
153
χ2=14.302 · df=1 · φ=0.319 · p=0.000

```r
apa.aov.table(one.way, filename="Table_anova.doc", table.number = 2)
```

```
##
##
## Table 2
##
## ANOVA results using sbp as the dependent variable
##
##
##    Predictor        SS  df        MS       F    p partial_eta2
##  (Intercept) 284760.89   1 284760.89 4990.82 .000
##     exercise   1064.03   2    532.01    9.32 .000          .11
##        Error   8558.54 150     57.06
##  CI_90_partial_eta2
##
##          [.04, .19]
##
##
## Note: Values in square brackets indicate the bounds of the 90% confidence
interval for partial eta-squared
```